View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

Harvard University

Harvard University Biostatistics Working Paper Series

Year 2007 *Paper* 64

Identifying patients who need additional biomarkers for better prediction of health outcome or diagnosis of clinical phenotype

Lu Tian* Tianxi Cai†

L. J. Wei‡

*Northwestern University, l-tian@northwestern.edu

†Harvard University, tcai@hsph.harvard.edu

‡Harvard University, wei@hsph.harvard.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/harvardbiostat/paper64

Copyright ©2007 by the authors.

Identifying patients who need additional biomarkers for better prediction of health outcome or diagnosis of clinical phenotype

LU TIAN

Department of Preventive Medicine,
Northwestern University, Chicago, IL 60611, USA

TIANXI CAI

Department of Biostatistics,
Harvard University, Boston, MA 02115, USA

LJ WEI*

Department of Biostatistics,
Harvard University, Boston, MA 02115, USA
wei@hsph.harvard.edu

SUMMARY

Suppose that we are interested in using new biomarkers to improve prediction or diagnosis of the patient's clinical phenotype in addition to the conventional markers. The incremental value from the new markers is typically assessed by averaging across patients in the entire population of interest. However, when measuring the new markers is costly or invasive, an overall improvement does not justify measuring the new markers in all patients. A more practical strategy is to utilize the patient's conventional markers to decide whether the new markers are needed for improving prediction of his/her health outcomes. In this article, we propose inference procedures for the incremental values of new markers across various subgroups of patients classified by the conventional markers. The resulting point and interval estimates can be quite useful for medical decision makers seeking to balance the predictive or diagnostic value of new markers against their associated cost and risk. Our proposals are theoretically justified and illustrated empirically with two real examples.

Keywords: Biomarker, Clinical Outcome, Diagnosis, Incremental Prediction Accuracy, K-fold Cross Validation, Prediction, Subgroup Analysis

1. INTRODUCTION

Biological and technological advances continually generate promising new biomarkers with the potential to improve medical care by providing more accurate, personalized predictions of health outcomes

*To whom correspondence should be addressed.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

and diagnoses of clinical phenotypes. However, extensive use of new markers may provide only negligible improvements in prediction or diagnosis, while subjecting patients to additional risks and costs. It is therefore important to develop statistical methods that can quantify for individual patients the value of new markers over conventional ones, especially when measuring these markers is costly or invasive. As an example, in a recent study, the incremental values from ten new biomarkers for prediction of first major cardiovascular events and death in a Framingham Study cohort were examined extensively (Wang *and others*, 2006). There were 3209 participants in the study. They were followed for a median of 7.4 years, during which 207 participants died and 169 had a first major cardiovascular event. Based on various prediction precision criteria, the study team found that the ten contemporary biomarkers added only moderate *overall* predictive value to the classical risk factors. In contrast, other investigators studying different populations with different prediction precision measures demonstrated that certain biomarkers provide clinically useful prognostic information on top of, for example, the traditional Framingham risk score for heart diseases (Ridker *and others*, 2002, 2007; Blumenthal *and others*, 2007).

Despite these often controversial findings in the literature, clinical practitioners would generally not change their recommendation for the patient's care with the extra marker information if the patient, for example, has either *high* or *very low* conventional risk score. Therefore, a practically important question is how to *systematically* identify patients who would benefit from the additional markers instead of evaluating these markers based only on their average incremental value across the entire population (D'Agostino , 2006). In this article, we propose procedures to estimate the incremental values of new markers for diagnosis or prediction in various subgroups of patients classified by conventional markers. These, coupled with the sampling variations of the estimates, provide a useful tool for researchers and practitioners to decide when, after observing the conventional risk factors, the new markers are needed. In Section 2, we describe in detail the new procedure and provide theoretical justification. In Section 3, we illustrate our methods with two examples, one with a continuous response and the other with a binary outcome.

There are quite a few procedures in the literature for evaluating the over-all incremental value of new markers for an entire population of interest. For example, Pepe *and others* (2004) compared the ROC curves among models with and without an additional marker. Recently, Tian *and others* (2007) and Uno *and others* (2007) proposed robust inference procedures for evaluating prediction rules. Prediction or diagnostic precision measures, which may be used for comparing different prediction procedures, have also been proposed and utilized, for example, by Brier (1950), Breiman *and others* (1984), Speigelhalter (1986), Korn and Simon (1990), McLachlan (1992), Mittlböck and Schemper (1996), Ripley (1996), Zhou *and others* (2002) and Pepe (2003).

2. ESTIMATING SUBJECT-SPECIFIC PREDICTION ERROR BASED ON RISK SCORE CONSTRUCTED FROM CONVENTIONAL MARKERS

Let Y be a continuous or binary response variable, U be the set of its conventional marker values, and V be the corresponding counterpart from the new markers. Our data consist of n independent copies $\{(Y_i, U_i, V_i), i = 1, \cdots, n\}$ from (Y, U, V). The problem is how to use the data to identify future subjects via U, which would benefit from the new markers for better prediction of their responses Y. Suppose that there are no well-established rules for classifying subjects based on U for predicting Y. First, we may estimate a center value of Y given U nonparametrically and use this estimate to construct a predictor for Y. We then estimate the average prediction error, the "distance" between the observed response and its predicted value over all subjects which have the same marker value U. Next, we estimate the center of Y given U and V, and estimate the corresponding average prediction error conditional only on U. Inferences about the improvement from the new markers can be made via these functional estimates over U. Unfortunately, in general, we can only construct nonparametric functional estimates, which behave

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

reasonably well, when the dimension of U is very small and the sample size n is quite large.

A practically feasible alternative to handle this problem is to consider a parametric or semi-parametric approach. To this end, let X be a p-dimensional vector, a function of U. Assume that the conditional mean of Y given U can be approximated by the following working model

$$E(Y \mid U) = g_1(\beta' X), \tag{2.1}$$

where $g_1(\cdot)$ is a smooth, strictly increasing, known function and β is an unknown vector of parameters. Note that the first component of X is one. In this article, we deal with the interesting and challenging case that $\beta' X$ is a continuous variable.

To estimate the regression parameters for model (2.1) which, most likely, is an approximation to the true conditional mean of Y given U, one may use the estimator $\widehat{\beta}$ based on the simple estimating function

$$S_1(\beta) = \sum_{i=1}^{n} X_i \left\{ Y_i - g_1(\beta' X_i) \right\}, \tag{2.2}$$

where $\{(Y_i, X_i), i = 1, \cdots, n\}$ are n-independent copies of (Y, X) (Tian *and others*, 2007). Note that even when (2.1) is not the true model, $\widehat{\beta}$ converges to a constant vector β_0, as $n \to \infty$. It is not clear, however, that other standard estimators for β in (2.1) would be convergent as n gets large.

Now, consider a future independent subject with $(Y, X) = (Y^0, X^0)$. For a given β in (2.1), let $\widehat{Y}_1(\beta' X^0)$ be the predictor for Y^0. For example, when Y^0 is continuous, one may let $\widehat{Y}_1(\beta' X^0) = g_1(\beta' X^0)$ and when Y^0 is binary, one may predict Y^0 by a binary variable $\widehat{Y}_1(\beta' X^0) = I\{g_1(\beta' X^0) \geqslant 0.5\}$, where $I(\cdot)$ is the indicator function. Other prediction rules for the binary case will be discussed in the Example Section. To evaluate the performance of $\widehat{Y}_1(\widehat{\beta}' X^0)$, we first need to quantify its prediction accuracy based on a "distance" between the true Y^0 and the predicted $\widehat{Y}_1(\widehat{\beta}' X^0)$, denoted by $D\{Y^0, \widehat{Y}_1(\widehat{\beta}' X^0)\}$. For example, one may let $D(a, b) = |a - b|$. For the binary case, this distance function is simply $I(a \neq b)$. Other choices of distance functions will be discussed in Section 3.

Next, since clinical practitioners almost always group subjects with a "risk scoring system" for medical decision making, we consider an average prediction error over a set of X's which have "similar" $g_1(\widehat{\beta}' X)$ to evaluate $\widehat{Y}_1(\cdot)$. To be specific, let $J_z = (c_z, d_z)$ be a data-independent interval centered about z, where z ranges over a set of possible values of $g_1(\beta_0' X)$. The average prediction error over J_z is $\mathcal{D}_1^{\star}(z) = E[D\{Y^0, \widehat{Y}_1(\widehat{\beta}' X^0)\} |\, g_1(\widehat{\beta}' X^0) \in J_z]$, where the conditional expectation is taken with respect to (Y^0, X^0) and $\widehat{\beta}$. As $n \to \infty$, $\mathcal{D}_1^{\star}(z)$ converges to

$$\mathcal{D}_1(z) = E\left[D\{Y^0, \widehat{Y}_1(\beta_0' X^0)\} |\, g_1(\beta_0' X^0) \in J_z \right], \tag{2.3}$$

where the expectation is taken with respect to (Y^0, X^0). As a process of z, this moving average process $\{\mathcal{D}_1(z)\}$ provides a performance profile of $\widehat{Y}_1(\cdot)$ over all possible values of $g_1(\beta_0' X)$. The choices of J_z are discussed via two examples in the next section.

Now, let W be a $q \times 1$ vector, a function of U and V. Assume that a working model for the conditional mean of Y given U and V is

$$\mathrm{E}(Y \mid U, V) = g_2(\theta' W), \tag{2.4}$$

where $g_2(\cdot)$ is a smooth, strictly increasing, known function and θ is an unknown vector of parameters. The first component of W is one. Again, we assume that $g_2(\theta' W)$ is a continuous variable. Let $\widehat{\theta}$ be the estimator for θ obtained from the following simple estimating function

$$S_2(\theta) = \sum_{i=1}^{n} W_i \left\{ Y_i - g_2(\theta' W_i) \right\}, \tag{2.5}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

where $W_i, i = 1, \cdots, n$, are n independent copies of W. Let θ_0 be the limit of $\widehat{\theta}$. Consider a future independent $(Y, X, W) = (Y^0, X^0, W^0)$. Let $\widehat{Y}_2(\theta' W^0)$ be the predictor constructed from (2.4) with parameter value θ, the counterpart of $\widehat{Y}_1(\beta' X^0)$. For the aforementioned interval J_z, let the average prediction error for $\widehat{Y}_2(\cdot)$ over J_z be

$$\mathcal{D}_2(z) = E\left[D\{Y^0, \widehat{Y}_2(\theta_0' W^0)\} \middle| \; g_1(\beta_0' X^0) \in J_z \right], \tag{2.6}$$

where the expectation is taken with respect to (Y^0, X^0, W^0). Then, as a process in z,

$$\Delta(z) = \mathcal{D}_1(z) - \mathcal{D}_2(z) \tag{2.7}$$

provides a global picture for identifying subgroups of patients who would benefit from the additional markers.

To estimate $\mathcal{D}_1(z)$ and $\mathcal{D}_2(z)$, one may use

$$\widehat{\mathcal{D}}_1(z) = \frac{\sum_{i=1}^n D\{Y_i, \widehat{Y}_1(\widehat{\beta}' X_i)\} I\{g_1(\widehat{\beta}' X_i) \in J_z\}}{\sum_{i=1}^n I\{g_1(\widehat{\beta}' X_i) \in J_z\}} \tag{2.8}$$

and

$$\widehat{\mathcal{D}}_2(z) = \frac{\sum_{i=1}^n D\{Y_i, \widehat{Y}_2(\widehat{\theta}' W_i)\} I\{g_1(\widehat{\beta}' X_i) \in J_z\}}{\sum_{i=1}^n I\{g_1(\widehat{\beta}' X_i) \in J_z\}}, \tag{2.9}$$

respectively. We then let $\widehat{\Delta}(z) = \widehat{\mathcal{D}}_1(z) - \widehat{\mathcal{D}}_2(z)$ to estimate $\Delta(z)$. In Appendix A, we show that with the distance function $D(a, b) = |a - b|$ or a function thereof, the above three estimators are uniformly consistent over an interval Ω consisting of all z's whose intervals J_z's are properly in the support of $g_1(\beta_0' X)$. Similar arguments may be used for cases with other distance functions.

To make further inferences about the added value from the new markers for predicting the response, in Appendix A, we show that the limiting distributions of the processes $\widehat{\mathcal{W}}_1(z) = n^{1/2}\{\widehat{\mathcal{D}}_1(z) - \mathcal{D}_1(z)\}$, $\widehat{\mathcal{W}}_2(z) = n^{1/2}\{\widehat{\mathcal{D}}_2(z) - \mathcal{D}_2(z)\}$ and $\widehat{\mathcal{W}}(z) = n^{1/2}\{\widehat{\Delta}(z) - \Delta(z)\}$, are the same as those of the Gaussian processes $\mathcal{W}_1^*(z), \mathcal{W}_2^*(z)$ and $\mathcal{W}^*(z)$, respectively, for $z \in \Omega$. Here, realizations from these three Gaussian processes (6.3), (6.4) and (6.5) given in Appendix A can be generated easily for any interval of z, where $\widehat{D}_1(z)$ and $\widehat{D}_2(z)$ are well-defined. In practice, one may not able to construct reasonably well-behaved interval estimators for $D_l(z), l = 1, 2$, for z is the tail parts of Ω. To this end, let $\widehat{\Omega}$ be a set of z such that $J_z \subset [\eta_1, \eta_2]$, where $n^{-1}\sum_{i=1}^n I\{g(\widehat{\beta}' X_i) \leqslant \eta_1\} > d_1$, $n^{-1}\sum_{i=1}^n I\{g(\widehat{\beta}' X_i) \geqslant \eta_2\} > d_2$, and d_1 and d_2 are given positive numbers. Then, with the above large sample approximations, for $z \in \widehat{\Omega}$, a $(1-\alpha)$, $0 < \alpha < 1$, point-wise confidence interval for $D_l(z), l = 1, 2$, is

$$\widehat{D}_l(z) \pm \xi_{\alpha/2} \sigma_{\mathcal{W}_l^*(z)}. \tag{2.10}$$

Here, $\sigma^2_{\mathcal{W}_l^*(z)}$ is the variance of the random variable $\mathcal{W}_l^*(z)$ and ξ_α is the upper 100αth percentage point of the standard normal. Furthermore, a $(1-\alpha)$ simultaneous confidence band for $\{D_l(z), z \in \widehat{\Omega}\}$ is

$$\widehat{D}_l(z) \pm \tau_{l\alpha} \sigma_{\mathcal{W}_l^*(z)}, \tag{2.11}$$

where

$$\text{pr}\left\{ \sup_{z \in \widehat{\Omega}} \left| \mathcal{W}_l^*(z) / \sigma_{\mathcal{W}_l^*(z)} \right| < \tau_{l\alpha} \right\} \geqslant 1 - \alpha.$$

To construct interval estimators for $\Delta(z)$, it is important to note that $\widehat{\Delta}(z)$ has a degenerate limiting distribution when $\widehat{Y}_1(\beta_0' X) = \widehat{Y}_2(\theta_0' W)$ for all $g_1(\beta_0' X) \in J_z$. Therefore, to obtain reasonable

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

interval estimators in practice, we consider the set $\widetilde{\Omega} \subset \widehat{\Omega}$ such that for $z \in \widetilde{\Omega}$, $\sum_{i=1}^{n} I\{\widehat{Y}_1(\widehat{\beta}'X_i) \neq \widehat{Y}_2(\widehat{\theta}'W_i), g_1(\widehat{\beta}'X_i) \in J_z\}/\sum_{i=1}^{n} I\{g_1(\widehat{\beta}'X_i) \in J_z\} > d_3$, where d_3 is a given positive number. Then, for $z \in \widetilde{\Omega}$, a $(1-\alpha)$, $0 < \alpha < 1$, point-wise confidence interval for $\Delta(z)$, is

$$\widehat{\Delta}(z) \pm \xi_{\alpha/2}\sigma_{\mathcal{W}^*(z)}. \tag{2.12}$$

Here, $\sigma^2_{\mathcal{W}^*(z)}$ is the variance of the random variable $\mathcal{W}^*(z)$. Moreover, a $(1-\alpha)$ simultaneous confidence band for $\{\Delta(z), z \in \widetilde{\Omega}\}$ is

$$\widehat{\Delta}(z) \pm \tau_\alpha \sigma_{\mathcal{W}^*(z)}, \tag{2.13}$$

where

$$\text{pr}\left\{\sup_{z\in\widetilde{\Omega}} \left|\mathcal{W}^*(z)/\sigma_{\mathcal{W}^*(z)}\right| < \tau_\alpha\right\} \geqslant 1-\alpha.$$

Note that for the case with a continuous response Y, $\widehat{\Omega} = \widetilde{\Omega}$.

Now, since we use the entire data set to estimate the parameters in (2.1) and (2.4) and also to estimate the average prediction errors (2.3) and (2.6), $\widehat{\mathcal{D}}_1(\cdot)$ and $\widehat{\mathcal{D}}_2(\cdot)$ may be significantly underestimated. To reduce such potential bias, one may consider the commonly used K-fold cross validation scheme. Specifically, we randomly split the data into K disjoint subsets of about equal size and label them as $\mathcal{I}_k, k = 1, \cdots, K$. For each k, we use all the observations, which are *not* in $\mathcal{I}_k$, to estimate parameters in (2.1) and (2.4) via estimating functions (2.2) and (2.5), and then use the observations in $\mathcal{I}_k$ to estimate prediction errors $\mathcal{D}_1(\cdot)$ and $\mathcal{D}_2(\cdot)$ with (2.8) and (2.9). Let the resulting estimators be denoted by $\widehat{\mathcal{D}}_{1k}(\cdot)$ and $\widehat{\mathcal{D}}_{2k}(\cdot)$, respectively. The cross validated estimators for $\mathcal{D}_1(\cdot), \mathcal{D}_2(\cdot)$ and $\Delta(\cdot)$ are $\widetilde{\mathcal{D}}_1(\cdot) = K^{-1}\sum_{k=1}^{K}\widehat{\mathcal{D}}_{1k}(\cdot)$, $\widetilde{\mathcal{D}}_2(\cdot) = K^{-1}\sum_{k=1}^{K}\widehat{\mathcal{D}}_{2k}(\cdot)$ and $\widetilde{\Delta}(\cdot) = \widetilde{\mathcal{D}}_1(\cdot) - \widetilde{\mathcal{D}}_2(\cdot)$, respectively. Again, these estimators are uniformly consistent if K is relatively small with respect to n.

In Appendix B, we show that for large n, the distributions of the processes $\widetilde{\mathcal{W}}_1(\cdot) = n^{1/2}\{\widetilde{\mathcal{D}}_1(\cdot) - \mathcal{D}_1(\cdot)\}$, $\widetilde{\mathcal{W}}_2(\cdot) = n^{1/2}\{\widetilde{\mathcal{D}}_2(\cdot) - \mathcal{D}_2(\cdot)\}$ and $\widetilde{\mathcal{W}}(\cdot) = n^{1/2}\{\widetilde{\Delta}(\cdot) - \Delta(\cdot)\}$ can also be approximated well by those of $\mathcal{W}_1^*(\cdot), \mathcal{W}_2^*(\cdot)$ and $\mathcal{W}^*(\cdot)$, respectively. Point-wise and simultaneous confidence intervals for $\mathcal{D}_1(\cdot)$, $\mathcal{D}_2(\cdot)$, and $\Delta(\cdot)$ can then be constructed based on the cross validated estimates and their large sample distributions accordingly.

3. EXAMPLES

We use two examples to illustrate the new proposals. The first example is from a clinical trial conducted by the AIDS Clinical Trials Group, ACTG 320 (Hammer *and others*, 1997). The study demonstrates that for various response endpoints, on average the three-drug combination therapy consisting of indinarvir, zidovudine and lamivudine, is much better than the two drug combination without indinarvir for treating HIV infected patients. Unfortunately, even with this potent combination, some patients may not respond to treatment, but suffer from non-trivial toxicity. Therefore, for future patients' management, it is important to have a reliable model for predicting patient's treatment responses based on certain "baseline" markers. A general conception is to use the baseline CD4 count and HIV-RNA, a measure of viral load, and the early changes of these two markers after initiation of therapy for treatment guidance (Demeter *and others*, 2001). For resource-limited regions, however, the cost of obtaining HIV-RNA is relatively expensive. Therefore, a challenging question is when we need RNA in addition to CD4 for better prediction of patient's response.

Recently Tian *and others* (2007) demonstrated that, on a population average sense, neither the baseline nor early RNA change (from baseline to week 8) would add a clinically meaningful value for predicting the long term change of CD4 (from baseline to Week 24), an important measure of the patient's immune response. Here, we try to locate a subgroup of patients, if any, who would benefit from the expensive

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

marker RNA. To this end, let the response Y be the change of CD4 cell counts from Week 0 to 24, let U consist of age, baseline CD4 and the early change in CD4, and let V consist of the baseline RNA and the early change in RNA. For our analysis, in Models (2.1) and (2.4), we let $X = (1, U')'$, $W = (1, U', V')'$, and $g_1(\cdot)$ and $g_2(\cdot)$ be the identity function. Also, we let $\widehat{Y}_1(\beta' X) = \beta' X$, $\widehat{Y}_2(\theta' W) = \theta' W$, $D(a, b) = |a - b|$ and interval J_z be $[z - 10, z + 10]$ for $z \in \widehat{\Omega} = [15, 165]$. In our analysis, we let $d_1 = d_2 = 0.01$ discussed in Section 2 for choosing $\widehat{\Omega}$. With $n = 392$ sets of complete observations of (Y, U, V), the regression parameter estimates for Models (2.1) and (2.4) are reported in Table 1. Note that the short term changes of CD4 and RNA are statistically highly significant.

Table 1. Estimates of the regression parameters with their standard errors and corresponding p-values for testing zero covariate effects for the AIDS example

	Age	Baseline RNA	RNA Change	Baseline CD4	CD4 Change
Estimate	-0.55	0.08	-12.06	0.03	0.68
Std Error	0.35	5.53	2.80	0.07	0.10
P-value	0.12	0.99	0.00	0.72	0.00

For both working models, we utilized 5-fold cross validation scheme discussed in Section 2 to obtain the regression parameters and then $\widetilde{\mathcal{D}}_1(\cdot)$, $\widetilde{\mathcal{D}}_2(\cdot)$, and $\widetilde{\Delta}(\cdot)$. In Figure 1, we present these estimated prediction errors and their differences with the corresponding 0.95 point-wise and simultaneous confidence intervals given in (2.10)-(2.13). The values of $\{\widetilde{\mathcal{D}}_1(z)\}$ based on the model with age, baseline CD4 and early change in CD4 range from 37 to 74. The values of $\{\widetilde{\mathcal{D}}_2(z)\}$ based on the model with additional RNA information range from 36 to 73. The estimated differences $\{\widetilde{\Delta}(z)\}$ range from -1.7 to 6.0. These indicate that there is no clinically meaningful gain from RNA for any subgroup of patients classified by $\widehat{\beta}' X$. One may draw further statistical inference about the $\Delta(\cdot)$. For example, for subjects whose score $g_1(\widehat{\beta}' X) \in J_z = [40, 60]$, the estimated $\widetilde{\Delta}(50) = 0.45$ with 0.95 point-wise interval of $(-3.25, 4.15)$ and simultaneous interval of $(-7.48, 8.38)$. Note that the results reported here are based on J_z with interval length of 20, which is well within the intra-patient variation of CD4 measures. Various analyses have also been done with J_z's whose lengths range from 30 to 60. All the results lead to the same conclusion. That is, statistically or clinically, we cannot identify a subgroup of patients who would benefit from the extra information of HIV-RNA for prediction of the long term CD4 change.

The data for the second example is from a population of patients screened for a clinical study, called TRACE, for treating heart failure or acute myocardial infraction (MI) (Kober *and others*, 1995). There were 6676 patients screened. Each patient had six routine clinical covariates: age, creatine (CRE), occurrence of heart failure (CHF), history of diabetes (DIA), history of hypertension (HYP), and cardiogenic shock after MI (KS). Moreover, each patient had an echocardiographic assessment of left ventricular systolic function which was quantified by a measure called the wall motion index (WMI). Compared with the above six covariates, the WMI is relatively expensive to obtain. Although not every screened patient entered the clinical trial, all patients screened were followed closely for mortality.

Recently, Thune *and others* (2005) studied the prognostic importance of left ventricular systolic function in patients diagnosed with either heart failure or acute MI in addition to the patient's medical history. It would be interesting to identify subpopulations that can benefit from the extra WMI measure for predicting clinical outcomes such as mortality. Here, we let the outcome Y be a binary variable, which is one if the patient died within five years. The five-year survival rate for this data set is approximately 42%. To evaluate the incremental value of WMI, we first fit the data using Model (2.1) with $X = (1, \text{AGE}, \text{CRE}, \text{CHF}, \text{DIA}, \text{HYP}, \text{KS})$, and $g_1(s) = \exp(s)/\{1 + \exp(s)\}$. With the extra variable

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

WMI, we fit a second logistic regression model with $W = (X', \text{WMI})'$. A total of 5921 subjects have complete predictor information. The estimates for the regression parameters with their standard errors are reported in Table 2. Note that the WMI is highly statistically significant.

Table 2. Estimated Regression Coefficients for Model (2.1) with AGE, CRE, CHF, DIA, HYP, KS and WMI for the screened population of TRACE study

Estimate	0.055	−0.010	0.759	0.718	0.187	1.153	−1.097
Std. Error	0.004	0.002	0.067	0.101	0.073	0.163	0.083
P-value	0.000	0.000	0.000	0.000	0.010	0.000	0.000

Next, we consider the prediction rules

$$\widehat{Y}_1(\beta' X) = I\{g_1(\beta' X) \geqslant c\}, \tag{3.1}$$

and

$$\widehat{Y}_2(\theta' W) = I\{g_2(\theta' W) \geqslant c\}. \tag{3.2}$$

Moreover, let $D(a, b) = |a - b|$. Now, for $c = 0.5$, the 5-fold cross validated estimates obtained by letting J_z be the entire real line in (2.8) and (2.9) for the overall prediction errors $E[D(Y^0, \widehat{Y}_1(\widehat{\beta}' X^0))]$ and $E[D(Y^0, \widehat{Y}_2(\widehat{\beta}' W^0))]$ are 0.28 and 0.26, respectively, a modest overall incremental gain from the extra information of WMI for the entire population of interest. To identify which subgroup of patients who would benefit with WMI, we let $J_z = [z - 0.1, z + 0.1]$, for $z \in \widehat{\Omega} = [0.15, 0.82]$. Here, $\widehat{\Omega}$ is chosen by letting $d_1 = d_2 = 0.01$ discussed in Section 2. To estimate $D_l(z), l = 1, 2$, and $\Delta(z)$, we used the 5-fold cross validation to obtain $\widetilde{\mathcal{D}}_1(\cdot)$, $\widetilde{\mathcal{D}}_2(\cdot)$ and $\widetilde{\Delta}(\cdot)$. In Figure 2, we present these point estimates and their corresponding 0.95 point-wise and simultaneous confidence intervals. For the interval estimation, we let $d_3 = 0.01$. This results in $\widetilde{\Omega} = [0.26, 0.76]$. Note that the point estimates $\widetilde{\Delta}(z)$ for z outside $\widehat{\Omega}$ are not reliable, and $\widetilde{\Delta}(z)$ is pretty flat around 0 for $z \in \widehat{\Omega} - \widetilde{\Omega}$, indicating that there is no evidence that WMI has a meaningful gain outside the interval $\widetilde{\Omega}$. On the other hand, with the point and interval estimates displayed in Figure 2(c), one may conclude that WMI is likely to be beneficial for patients with conventional risk scores $g_1(\widehat{\beta}' X)$ ranging from 0.16 to 0.74. If WMI is relatively affordable to the population of interest, then one may consider using the upper bound of the simultaneous confidence intervals to identify the subpopulation based on $\widetilde{\Delta}(z) + \tau_\alpha \sigma_{\mathcal{W}^*(z)} \geqslant 0$ and thus conclude that patients with $g_1(\widehat{\beta}' X) \in [0.16, 0.86]$ are likely to benefit from the WMI. On the other hand, when WMI is not quite affordable, then one may select the region conservatively and use the lower bound of the simultaneous confidence intervals based on $\widetilde{\Delta}(z) - \tau_\alpha \sigma_{\mathcal{W}^*(z)} \geqslant 0$ and thus conclude that patients with $g_1(\widehat{\beta}' X) \in [0.29, 0.63]$ are likely to benefit from the WMI.

Note that for any prediction rule $\widehat{Y}$, the conditional or unconditional expectation of the above distance function $D(Y, \widehat{Y})$ consists of two discordance rates or two types of error rates. For example, $\mathcal{D}_1(z)$ in (2.3) is $\mathcal{D}_{11}(z) + \mathcal{D}_{10}(z)$, where $\mathcal{D}_{11}(z) = E[Y^0 D\{1, \widehat{Y}_1(\beta_0' X^0)\} |\ g_1(\beta_0' X^0) \in J_z]$, the discordance rate for false negative errors, and $\mathcal{D}_{10}(z) = E[(1 - Y^0) D\{0, \widehat{Y}_1(\beta_0' X^0)\} |\ g_1(\beta_0' X^0) \in J_z]$, the discordance rate for false positive errors. The $\mathcal{D}_{20}(z)$ and $\mathcal{D}_{21}(z)$ are similarly defined. Let $\Delta_0(z) = \mathcal{D}_{10}(z) - \mathcal{D}_{20}(z)$ and $\Delta_1(z) = \mathcal{D}_{11}(z) - \mathcal{D}_{21}(z)$. Oftentimes, a false negative conclusion may lead to a more serious consequence than a false positive. Therefore, one may consider a weighted sum of $\Delta_0(z)$ and $\Delta_1(z)$, $\Delta(w, z) = w_0 \Delta_0(z) + w_1 \Delta_1(z)$, to evaluate the importance of the extra markers, where $w = (w_0, w_1)'$ and w_0 and w_1 are non-negative constants. For a given w, the cross validated point estimates $\widetilde{\Delta}(w, z)$ and their interval estimates for $\Delta(w, z)$ can be constructed as for $\Delta(z)$ in Section 2.

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

In Figure 3(a),(b),(c) and (d), we present the point and interval estimates of $\Delta(w, z)$ for the predictors (3.1) and (3.2) with $c = 0.5$ and various choices of w. Note that when $w_0 \neq w_1$, even if the working model is correctly specified, the prediction rule in (3.1) or (3.2) with $c = 0.5$ is not optimal with respect to the weighted error rate. Furthermore, with the unequal weighting criterion, for some subgroups of patients, inclusion of the extra information of WMI may significantly decrease the prediction precision.

For a given w, with the weighted sum prediction precision measure, $w_0\mathcal{D}_{10}(z) + w_1\mathcal{D}_{11}(z)$, it is straightforward to show that the optimal prediction rule based on X that minimizes the above criterion is $\widehat{Y} = I\{\text{pr}(Y = 1 \mid X) \geqslant c_w\}$, where $c_w = w_0/(w_0 + w_1)$. Therefore, for the present example, if $g_1(\widehat{\beta}'X)$ and $g_2(\widehat{\theta}'W)$ are reasonably good approximations to $E(Y|U)$ and $E(Y|U, V)$, the predictors $I(g_1(\widehat{\beta}'X) \geqslant c_w)$ and $I(g_2(\widehat{\theta}'W) \geqslant c_w)$ are almost optimal. In Figure 4, we present the cross validated point estimates along with the 0.95 interval estimates of $\Delta(w, z)$ with $w = (1, 4)'$ and $(1, 9)'$ when "optimal" prediction rules are used for both models. It appears that there is minimal gain from WMI across all sub-populations indexed by $g(\widehat{\beta}'X) \in J_z$ for both cases.

4. REMARKS

From the results of our analysis presented in the Example Section, we find that the decision to include or exclude the additional biomarkers for prediction of a patients' health outcome depends heavily on the prediction precision measure or utility function. In the cardiovascular disease arena, clinicians may recommend certain treatments to patients whose predicted 10-year risk of having a cardiovascular event is higher than, for example, 10%. The utility or cost function for choosing this cutoff points can be rather complex, if not impossible, to quantify. Furthermore, the utility function may vary across individuals and hence different patients may have different optimal cutoff points for predicting patient-level outcomes. The weighted sum of prediction error rates presented in this article is an attempt to cope with this complicated cost-benefit issue. The complexities of choosing a loss function extend to the case of continuous responses. For example, weighting absolute prediction errors according to the observed response may lead to a more meaningful penalty in some cases than the un-weighted absolute prediction error.

The proposed methods may be extended to the case where responses are event times subjected to censoring. Since the support of the censoring variable is usually shorter than that of the event time in practice, we may utilize the approach taken by Uno *and others* (2007) and construct predictors for t-year survival. It would be interesting to investigate whether the additional biomarkers are useful for predicting long- or short-term survivors, with potentially different subsets of patients benefiting in each case.

5. ACKNOWLEDGEMENT

The authors are grateful to Dr. James Signorovitch for helpful comments on the paper. This research was partially supported by the US NIH grants for I2B2 and AIDS

6. APPENDIX A

Large sample properties of $\widehat{\mathcal{D}}_1(\cdot)$, $\widehat{\mathcal{D}}_2(\cdot)$ and $\widehat{\Delta}(\cdot)$

To justify the asymptotic properties of the proposed estimators, certain *smooth* regularity conditions are needed for the distance function $D(\cdot, \cdot)$ and its corresponding predictor. Here, we consider the case that the distance function is $D(Y, \widehat{Y}) = |Y - \widehat{Y}|$ for continuous and $w_0^{(1-Y)}w_1^Y|Y - \widehat{Y}|$ for binary responses, where w_0 and w_1 are given positive numbers. Furthermore, when Y is continuous, we let $\widehat{Y}_1(\beta'x) = g_1(\beta'x)$, and $\widehat{Y}_2(\theta'w) = g_2(\theta'w)$, and when Y is binary, let $\widehat{Y}_1(\beta'x) = I\{g_1(\beta'x) \geqslant \text{constant}\}$, and

BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

$\widehat{Y}_2(\theta' w) = I\{g_2(\theta' w) \geqslant \text{constant}\}$. Similar arguments can be used to justify other cases.

Suppose that β_0 and θ_0 are interior points of their compact parameter spaces. Let Ω denote the set of z such that J_z is properly contained in the support of $\beta_0' X$. First, we show that the above estimators are uniformly consistent over Ω. To this end, let $\widehat{\Theta} = (\widehat{\beta}', \widehat{\theta}')'$ and $\Theta = (\beta', \theta')'$

$$\widehat{\mathcal{D}}_1(z, \beta) = \frac{\sum_{i=1}^n D\{Y_i, \widehat{Y}_1(\beta' X_i)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i=1}^n I\{g_1(\beta' X_i) \in J_z\}},$$

$$\widehat{\mathcal{D}}_2(z, \Theta) = \frac{\sum_{i=1}^n D\{Y_i, \widehat{Y}_2(W_i, \theta)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i=1}^n I\{g_1(\beta' X_i) \in J_z\}},$$

$\mathcal{D}_1(z, \beta) = E[D\{Y, \widehat{Y}_1(\beta' X)\} | g_1(\beta' X) \in J_z]$ and $\mathcal{D}_2(z, \Theta) = E[D\{Y, \widehat{Y}_2(\theta' W)\} | g_1(\beta' X) \in J_z]$. It follows from the uniform law of large numbers (Pollard, 1990, Ch. 8) that $\sup_{z,\beta} |\widehat{\mathcal{D}}_1(z, \beta) - \mathcal{D}_1(z, \beta)| + \sup_{z,\Theta} |\widehat{\mathcal{D}}_2(z, \Theta) - \mathcal{D}_2(z, \Theta)|$ converges to 0, in probability, where the sup is taken over Ω and the parameter spaces. This, together with the convergence property of $\widehat{\beta}$ and $\widehat{\theta}$, implies the uniform consistency of $\widehat{\mathcal{D}}_1(z) = \widehat{\mathcal{D}}_1(z, \widehat{\beta})$ and $\widehat{\mathcal{D}}_2(z) = \widehat{\mathcal{D}}_2(z, \widehat{\Theta})$. The consistency of $\widehat{\Delta}(\cdot)$ follows accordingly.

Next, we show that the processes $\widehat{\mathcal{D}}_1(\cdot)$, $\widehat{\mathcal{D}}_2(\cdot)$ and $\widehat{\Delta}(\cdot)$ after standardization are asymptotically normal. First, let $T = (Y, U', V')'$ and $\widehat{\Theta} = (\widehat{\beta}', \widehat{\theta}')'$. It follows from Appendix 1 of Tian *and others* (2007), $n^{\frac{1}{2}}(\widehat{\Theta} - \Theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n \psi(T_i) + o_p(1)$, where $\psi(T) = \{\psi_1(T)', \psi_2(T)'\}'$,

$$\psi_1(T) = [E\{\dot{g}_1(\beta_0' X) X X'\}]^{-1} X\{Y - g_1(\beta_0' X)\}, \;\; \psi_2(T) = [E\{\dot{g}_2(\theta_0' W) W W'\}]^{-1} W\{Y - g_2(\theta_0' W)\},$$

and $\dot{g}_k(\cdot)$ is the derivative of $g_k(\cdot)$. Now, let $\widehat{\mathcal{W}}_1(z, \beta) = n^{\frac{1}{2}} \left\{\widehat{\mathcal{D}}_1(z, \beta) - \mathcal{D}_1(z, \beta)\right\}$ and $\widehat{\mathcal{W}}_2(z, \Theta) = n^{\frac{1}{2}} \left\{\widehat{\mathcal{D}}_2(z, \Theta) - \mathcal{D}_2(z, \Theta)\right\}$. By the maximum inequality for the standard empirical processes (Pollard, 1990, Ch.9),

$$\sup_{z,\beta} \left| \widehat{\mathcal{W}}_1(z, \beta) - n^{-\frac{1}{2}} \sum_{i=1}^n \xi_1(z, \beta, T_i) \right| + \sup_{z,\Theta} \left| \widehat{\mathcal{W}}_2(z, \Theta) - n^{-\frac{1}{2}} \sum_{i=1}^n \xi_2(z, \Theta, T_i) \right| \to 0,$$

in probability, as $n \to \infty$, where

$$\xi_1(z, \beta, T_i) = \frac{I\{g_1(\beta' X_i) \in J_z\} \left[D\{Y_i, \widehat{Y}_1(\beta' X_i)\} - \mathcal{D}_1(z, \beta)\right]}{\text{pr}(g_1(\beta' X) \in J_z)},$$

and

$$\xi_2(z, \Theta, T_i) = \frac{I\{g_1(\beta' X_i) \in J_z\} \left[D\{Y_i, \widehat{Y}_2(W_i, \theta)\} - \mathcal{D}_2(z, \Theta)\right]}{\text{pr}(g_1(\beta' X) \in J_z)}.$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

This, together with the above linear expansion of $n^{\frac{1}{2}}(\widehat{\Theta}-\Theta_0)$, implies that

$$\begin{aligned}\widehat{\mathcal{W}}_1(z) &= n^{\frac{1}{2}}\left\{\widehat{\mathcal{D}}_1(z,\widehat{\beta})-\mathcal{D}_1(z,\widehat{\beta})+\mathcal{D}_1(z,\widehat{\beta})-\mathcal{D}_1(z,\beta_0)\right\}\\ &\simeq n^{\frac{1}{2}}\left\{\widehat{\mathcal{D}}_1(z,\beta_0)-\mathcal{D}_1(z,\beta_0)\right\}+\dot{\mathcal{D}}_1^{(2)}(z,\beta_0)'n^{\frac{1}{2}}(\widehat{\beta}-\beta_0)\\ &\simeq n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{\xi_1(z,\beta_0,T_i)+\dot{\mathcal{D}}_1^{(2)}(z,\beta_0)'\psi_1(T_i)\right\} \qquad (6.1)\\ \widehat{\mathcal{W}}_2(z) &= n^{\frac{1}{2}}\left\{\widehat{\mathcal{D}}_2(z,\widehat{\Theta})-\mathcal{D}_2(z,\widehat{\Theta})+\mathcal{D}_2(z,\widehat{\Theta})-\mathcal{D}_2(z,\Theta_0)\right\}\\ &\simeq n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{\xi_2(z,\Theta_0,T_i)+\dot{\mathcal{D}}_2^{(2)}(z,\Theta_0)'\psi(T_i)\right\} \qquad (6.2)\end{aligned}$$

where $\dot{\mathcal{D}}_1^{(2)}(z,\beta)=\partial\mathcal{D}_1(z,\beta)/\partial\beta$ and $\dot{\mathcal{D}}_2^{(2)}(z,\Theta)=\partial\mathcal{D}_2(z,\Theta)/\partial\Theta$. It follows from a functional central limit theorem (Pollard, 1990, Ch. 10) that the processes $\widehat{\mathcal{W}}_1(\cdot)$ and $\widehat{\mathcal{W}}_2(\cdot)$ converge weakly to zero-mean Gaussian processes. The weak convergence of $\widehat{\mathcal{W}}(\cdot)$ follows accordingly.

To approximate the distribution of the processes $\widehat{\mathcal{W}}_1(\cdot)$, $\widehat{\mathcal{W}}_2(\cdot)$ and $\widehat{\mathcal{W}}(\cdot)$, we consider the perturbed version of these processes. The resulting processes are

$$\mathcal{W}_1^*(z)=n^{\frac{1}{2}}\left\{\frac{\sum_{i=1}^n[D\{Y_i,\widehat{Y}_1(X_i,\widehat{\beta})\}-\widehat{\mathcal{D}}_1(z,\widehat{\beta})]I\{g(\widehat{\beta}'X_i)\in J_z\}G_i}{\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\in J_z\}}+\widehat{\mathcal{D}}_1(z,\widehat{\beta}^*)-\widehat{\mathcal{D}}_1(z,\widehat{\beta})\right\}, \qquad (6.3)$$

$$\mathcal{W}_2^*(z)=n^{\frac{1}{2}}\left\{\frac{\sum_{i=1}^n[D\{Y_i,\widehat{Y}_2(W_i,\widehat{\theta})\}-\widehat{\mathcal{D}}_2(z,\widehat{\Theta})]I\{g(\widehat{\beta}'X_i)\in J_z\}G_i}{\sum_{i=1}^n I\{g(\widehat{\beta}'X_i)\in J_z\}}+\widehat{\mathcal{D}}_2(z,\widehat{\Theta}^*)-\widehat{\mathcal{D}}_2(z,\widehat{\Theta})\right\}, \qquad (6.4)$$

and

$$\mathcal{W}^*(z)=\mathcal{W}_1^*(z)-\mathcal{W}_2^*(z), \qquad (6.5)$$

where $\{G_1,...,G_n\}$ are independent standard normal random variables that are independent of the data, $\widehat{\Theta}^*=(\widehat{\beta}^{*'},\widehat{\theta}^{*'})'$,

$$\widehat{\beta}^*=\widehat{\beta}+\left\{\sum_{i=1}^n\dot{g}_1(\widehat{\beta}'X_i)X_iX_i'\right\}^{-1}\sum_{i=1}^n X_i\left\{Y_i-g_1(\widehat{\beta}'X_i)\right\}G_i$$

$$\text{and}\quad \widehat{\theta}^*=\widehat{\theta}+\left\{\sum_{i=1}^n\dot{g}_2(\widehat{\theta}'W_i)W_iW_i'\right\}^{-1}\sum_{i=1}^n W_i\left\{Y_i-g_2(\widehat{\theta}'W_i)\right\}G_i.$$

It follows from the same arguments as given above and similar arguments as in Appendix 4 of Cai *and others* (2005) that the limiting distributions of $\mathcal{W}_1^*(\cdot)$, $\mathcal{W}_2^*(\cdot)$ and $\mathcal{W}^*(\cdot)$, conditional on the data, are the same as those of $\widehat{\mathcal{W}}_1(\cdot)$, $\widehat{\mathcal{W}}_2(\cdot)$ and $\widehat{\mathcal{W}}(\cdot)$, respectively, on Ω. Since $\text{pr}(\widehat{\Omega}\subset\Omega)\to 1$, the confidence interval given in (2.10)is asymptotically valid for any $z\in\widehat{\Omega}$. Furthermore, noting the fact that $\sup_{\widehat{\Omega}}|\mathcal{W}_l^*(z)/\sigma_{\mathcal{W}_l^*(z)}|$ and $\sup_{\widehat{\Omega}}|\widehat{\mathcal{W}}_l(z)/\sigma_{\widehat{\mathcal{W}}_l(z)}|$ are asymptotically equivalent to $\sup_{\Omega_{d_1,d_2}}|\mathcal{W}_l^*(z)/\sigma_{\mathcal{W}_l^*(z)}|$ and $\sup_{\Omega_{d_1,d_2}}|\widehat{\mathcal{W}}_l(z)/\sigma_{\widehat{\mathcal{W}}_l(z)}|$, respectively, where $\Omega_{d_1,d_2}\subset\Omega$ is the limit of $\widehat{\Omega}$, the asymptotical confidence band over the random region $\widehat{\Omega}$ given in (2.11) is valid as well. Similarly, one may justify the

BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

validity of the confidence interval and band given in (2.12) and (2.13) by noting that $\widetilde{\Omega}_{d_3}$, the limit of $\widetilde{\Omega}$, is a subset of Ω and $\sigma_{\mathcal{W}(z)}$ is uniformly bounded below by a positive constant for $z \in \widetilde{\Omega}$.

7. APPENDIX B

Large sample properties of crossvalidated estimators

For each partition $\mathcal{I}_k$, let $\widehat{\Theta}_{(-k)} = (\widehat{\beta}'_{(-k)}, \widehat{\theta}'_{(-k)})'$ be the estimated Θ using data not in $\mathcal{I}_k$ via (2.2) and (2.5),

$$\widehat{\mathcal{D}}_{1k}(z,\beta) = \frac{\sum_{i\in\mathcal{I}_k} D\{Y_i, \widehat{Y}_1(\beta' X_i)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i\in\mathcal{I}_k} I\{g_1(\beta' X_i) \in J_z\}},$$

and

$$\widehat{\mathcal{D}}_{2k}(z,\Theta) = \frac{\sum_{i\in\mathcal{I}_k} D\{Y_i, \widehat{Y}_2(\theta' W_i)\} I\{g_1(\beta' X_i) \in J_z\}}{\sum_{i\in\mathcal{I}_k} I\{g_1(\beta' X_i) \in J_z\}}.$$

Since K is small with respect to n, $\widehat{D}_{1k}(z, \widehat{\beta}_{(-k)})$ is consistent. Then, it follows from the same argument in Appendix A, $n^{1/2}\{\widehat{D}_{1k}(z, \widehat{\beta}_{(-k)}) - \mathcal{D}_1(z,\beta_0)\}$ is asymptotically equivalent to

$$n^{-\frac{1}{2}} K \sum_{i=1}^{n} I(\tau_i = k)\xi_1(z,\beta_0,T_i) + n^{\frac{1}{2}}\dot{\mathcal{D}}_1(z,\beta_0)(\widehat{\beta}_{(-k)} - \beta_0),$$

where $\{\tau_i; i = 1, \cdots, n\}$ are n exchangeable discrete random variables uniformly distributed over $\{1, 2, \cdots, K\}$, independent of the data, and $\sum_{i=1}^{n} I(\tau_i = k) \approx n/K, k = 1, \cdots, K$. It follows from the same argument in Appendix 3 of Tian *and others* (2007) that conditional on the observed $\{\tau_i, i = 1, \cdots, n\}$

$$\widehat{\beta}_{(-k)} - \beta_0 = \frac{K}{n(K-1)} \sum_{i=1}^{n} I(\tau_i \neq k)\psi_1(T_i) + o_p(n^{-1/2}).$$

Then using the same argument in Appendix A, one can show that

$$\widetilde{\mathcal{W}}_1(z) = \frac{n^{\frac{1}{2}}}{K} \sum_{k=1}^{K} \left\{\widetilde{\mathcal{D}}_{1k}(z) - \mathcal{D}_1(z)\right\} = \frac{n^{-\frac{1}{2}}}{K} \sum_{i=1}^{n}\sum_{k=1}^{K} \left\{ I(\tau_i = k) K\xi_1(z,\beta_0,T_i) + \frac{KI(\tau_i \neq k)\psi_1(T_i)}{K-1} \right\}.$$

Since $\sum_{k=1}^{K} I(\tau_i = k) = 1$ and $\sum_{k=1}^{K} I(\tau_i \neq k) = K - 1$, it is straightforward to show that $\widetilde{\mathcal{W}}_1(z)$ is asymptotically equivalent to $\widehat{\mathcal{W}}_1(z)$ and thus the distribution of $\widetilde{\mathcal{W}}_1(\cdot)$ can be approximated by that of $\mathcal{W}_1^*(\cdot)$ conditional on the partition indicators $\{\tau_i, i = 1, \cdots, n\}$. Similar arguments can be used to show that the distributions of $\widetilde{\mathcal{W}}_2(\cdot)$ and $\widetilde{\mathcal{W}}(\cdot)$ can be approximated by those of $\mathcal{W}_2^*(\cdot)$ and $\mathcal{W}^*(\cdot)$, respectively.

REFERENCES

BLUMENTHAL, R., MICHOS, E. AND NASIR, K. (2007). Further improvements in CHD risk prediction for women. *JAMA* **297**, 641–43.

BREIMAN, L.,FRIEDMAN, J., STONE, C. AND OLSHEN, R. (1984). Classification and regression trees. Chapman & Hall/CRC.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

BRIER, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**, 1–3.

CAI, T., TIAN, L., AND WEI, L. J. (2005). Semiparametric Box-cox power transformation models for cnesored survival observations. *Biometrika* **92**, 619–32.

D'AGOSTINO, R. B. (2006). Risk prediction and finding new independent prognostic factors. *Journal of Hypertension* **24**, 643–45.

DEMETER, L., HUGHES, M., COOMBS, R., JACKSON, J., GRIMES, J., BOSCH, R., FISCUS, S., SPECTOR, S., SQUIRES, K., FISCHL, M., AND HAMMER, S. (2001). Predictors of virologic and clinical outcomes in HIV-1-infected patients receiving concurrent treatment with indinavir, zidovudine, and lamivudine. AIDS Clinical Trials Group Protocol 320 *Annals of Internal Medicine* **135**, 954–64.

HAMMER. S., SQUIRES, K., HUGHES, M., GRIMES, J., DEMETER, L., CURRIER, J., ERON, J., FEINBERG, J., BALFOUR, H., DEYTON, L., CHODAKEWITZ, J., AND FISCHL, M. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* **337**, 725–33.

KOBER, L., TORP-PEDERSEN, C., CARLSEN, J., BAGGER, H., ELIASEN, P., LYNGBORG, K., VIDEBAK, J., COLE, D., AUCLERT, L., PAULY, N., ALIOT, E., PERSSON, S., AND CAMM, A. (1995). A clinical trial of the angiotensin-converting-enzyme inhibitor trandolapril in patients with left ventricular dysfunction after myocardial infarction. *New England Journal of Medicine* **333**, 1670–76.

KORN, E. AND SIMON, R. (1990). Measures of explained variation for survival data. *Statistics in Medicine* **9**, 487–503.

MCLACHLAN, J. (1992). Discriminant analysis and statistical pattern recognition. John Wiley & Sons.

MITTLBÖCK, M. AND SCHEMPER, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine* **15**, 1987–97.

PEPE, M. S. (1992). The statistical evaluation of medical tests for classification and prediction. Oxford University Press.

PEPE, M.S., JANES, H., LONGTON, G., LEISENRING, W. AND NEWCOMB, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology* **159**, 882–90.

POLLARD, D. (1990). Empirical Processes: Theory and Applications. NSF-CMBS Regional Conference Series in Probability and Statistics 2. Hayward, CA: Institute of Mathematical Statistics.

RIDKER, P., RIFAI, N., ROSE, L., BURING, J. AND COOK, N. (2002). Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. *New England Journal of Medicine* **347**, 1557–65.

RIDKER, P., BURING, J., RIFAI, N. AND COOK, N. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds risk score. *JAMA* **297**, 611–19.

RIPLEY, B. (1996). Pattern recognition and neural networks. Cambridge University Press.

SPEIGELHALTER, D. (1986). Probabilistic predictin in patient management and clinical trials. *Statistics in Medicine* **5**, 421–33.

THUNE, J., CARLSEN, C., BUCH, P., SEIBAK, M., BURCHARDTB, H., TORP-PEDERSEN, C., AND KOBER L. (2005). Different prognostic impact of systolic function in patients with heart failure and/or acute myocardial infarction. *European Journal of Heart Failure* **7**, 852–8.

TIAN, L., CAI, T., GOETGHEBEUR, E., AND WEI, L. J. (2007). Model evaluation based on the distribution of estimated absolute prediction error. *Biometrika* Biometrika Advance Access, doi:10.1093/biomet/asm036.

UNO, H., CAI, T., TIAN, L. AND WEI, L. J. (2007). Evaluating prediction rules for t-Year survivors with censored regression models *Journal of American Statistical Association* **102**, 527-37.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

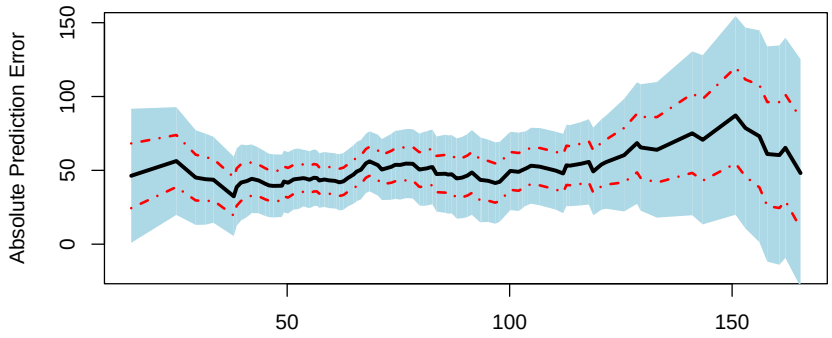
WANG, T., GONA, P., LARSON, M., TOFLER, G., LEVY, D., NEWTON-CHEH, C., JACQUES, P., RIFAI, N., SELHUB, J., ROBINS, S., BENJAMIN, E., D'AGOSTINO, R. AND VASAN, R. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death. *New England Journal of Medicine* **355**, 2631–39.

ZHOU, X. H., OBUCHOWSKI, N. A. AND MCCLISH, D. K. (2002). Statistical methods in diagnostic medicine. John Wiley & Sons.

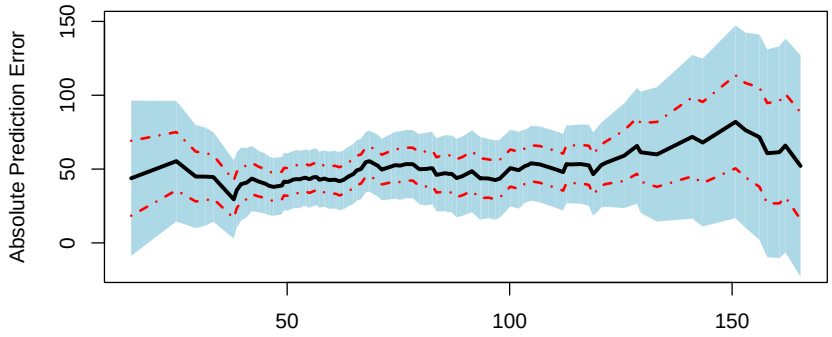
COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

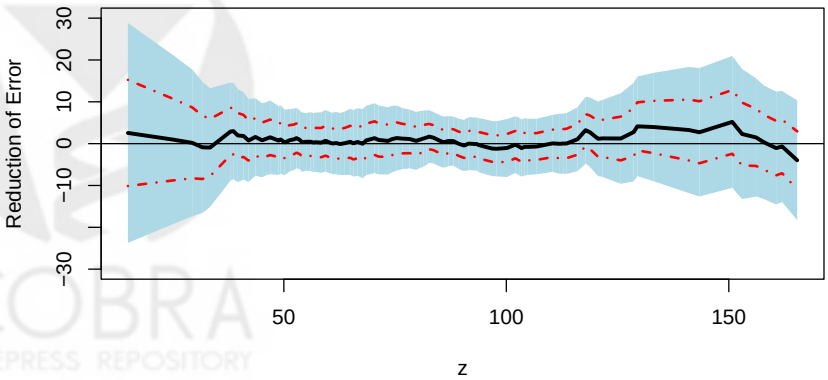
Fig. 1. Point estimates for $\mathcal{D}_1(\cdot)$, $\mathcal{D}_2(\cdot)$ and $\Delta(\cdot)$ with corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the HIV example.



(a) $\mathcal{D}_1(z)$, without HIV-RNA



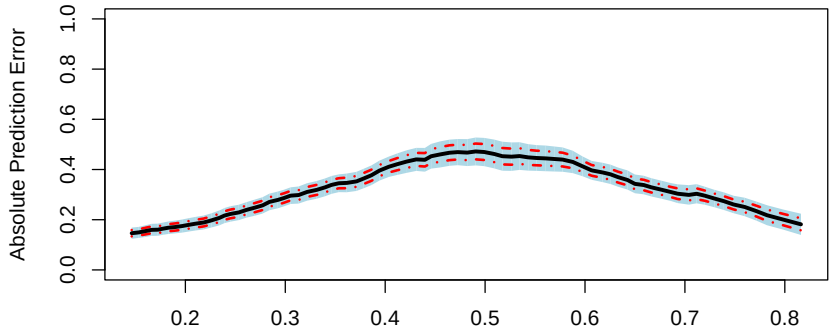
(b) $\mathcal{D}_2(z)$, with HIV-RNA



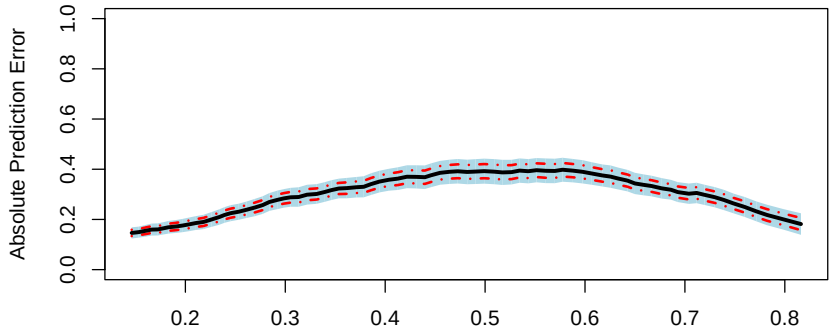
(c) $\Delta(z)$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

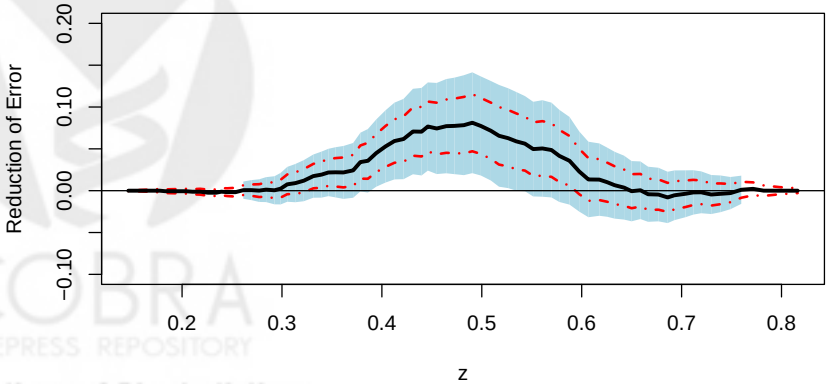
Fig. 2. Point estimates for $\mathcal{D}_1(\cdot)$, $\mathcal{D}_2(\cdot)$ and $\Delta(\cdot)$ with corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study (the prediction with $c = 0.5$).



(a) $\mathcal{D}_1(z)$, without WMI



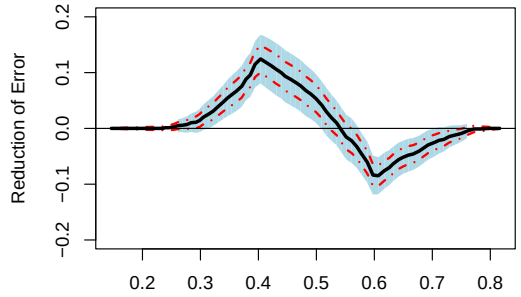
(b) $\mathcal{D}_2(z)$, with WMI



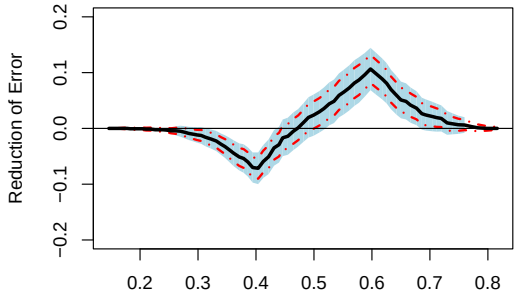
(c) $\Delta(z)$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

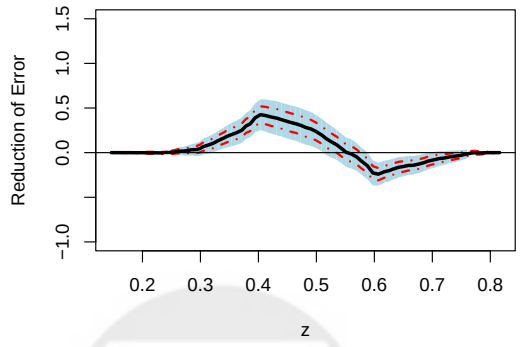
Fig. 3. Point estimate $\widetilde{\Delta}(w, \cdot)$ for $\Delta(w, \cdot)$ with various weights and the corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study (the prediction with $c = 0.5$).



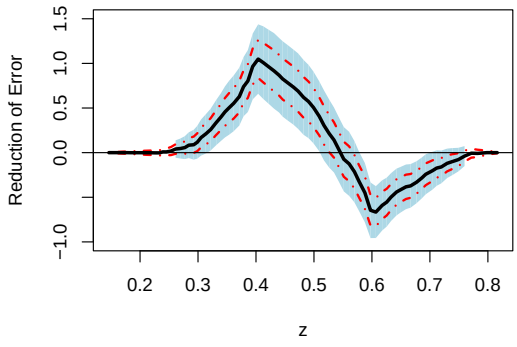
(a) $w_0 = 0, w_1 = 1$



(b) $w_0 = 1, w_1 = 0$



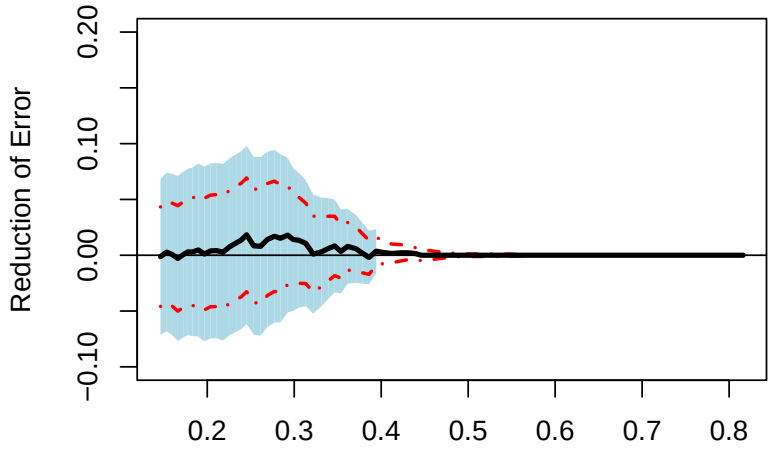
(c) $w_0 = 1, w_1 = 4$



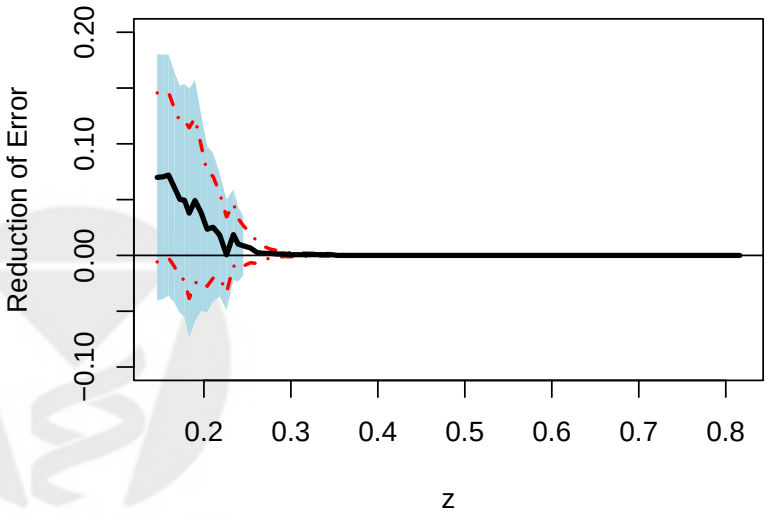
(d) $w_0 = 1, w_1 = 9$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Fig. 4. Point estimate $\widetilde{\Delta}(w, \cdot)$ for $\Delta(w, \cdot)$ with the "optimal" weights and the corresponding 0.95 point-wise (dashed lines) and simultaneous (shaded regions) confidence intervals for the screened population of the TRACE study.



(a) $w_0 = 1, w_1 = 4, c = 0.2$



(b) $w_0 = 1, w_1 = 9, c = 0.1$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive